Applications of Artificial Intelligence for Chemical Inference. III. Aliphatic Ethers Diagnosed by Their Low-Resolution Mass Spectra and Nuclear Magnetic Resonance Data[2]

Gustav Schroll,[3] A. M. Duffield, Carl Djerassi, B. G. Buchanan, G. L. Sutherland, E. A. Feigenbaum, and J. Lederberg

Contribution from the Departments of Chemistry, Computer Science, and Genetics, Stanford University, Stanford, California 94305. *Received May 16, 1969*

Abstract: A computer program capable of interpreting the low-resolution mass spectra of aliphatic ethers is described. This program (Heuristic DENDRAL) makes extensive use of the DENDRAL algorithm. In order to obtain unequivocal answers to the identity of unknown compounds an nmr subroutine was incorporated into the final stage of Heuristic DENDRAL. As demonstrated in Table I, the use of this program either resulted in one or two candidates, if the number of theoretical possibilities is relatively low (less than 200), or at least in a drastic reduction of the number of possible structures (*e.g.*, 10 candidates out of 989 possibilities) using solely mass spectral input. If nmr data are also employed, then a further reduction, usually leading to a single structure, is possible.

Although in its infancy, computer interpretation of both low-[1,4] and high-resolution[5,6] mass spectra shows considerable promise for the future automatic evaluation of experimental data. Our basic approach to this problem was enunciated in our earlier publications[1,7] where we described a computer program (Heuristic DENDRAL) capable of interpreting the low-resolution mass spectra of aliphatic ketones. This system relied extensively on the computer program DENDRAL which generates[7] complete and irredundant lists of aliphatic structures corresponding to any empirical composition. We have now extended the scope of Heuristic DENDRAL to include aliphatic ethers. This was a logical development in view of the well-documented[8] behavior of this class of compound in the mass spectrometer.

A diagrammatic representation of Heuristic DENDRAL is depicted in Scheme I. Given an unknown mass spectrum (Figure 1) and the empirical formula of the molecular ion the program must infer the presence of the correct functional group which in the present instance is the ether group. This information is obtained by the PRELIMINARY INFERENCE MAKER[9] and is then used by the STRUCTURE GENERATOR to compile exhaustive and irredundant lists of candidate structures containing this functional group. Truncation of the list of candidate structures is achieved by the PREDICTOR section of Heuristic DENDRAL in which a predicted mass spectrum

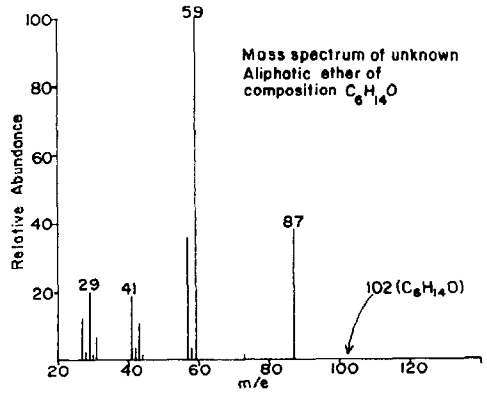
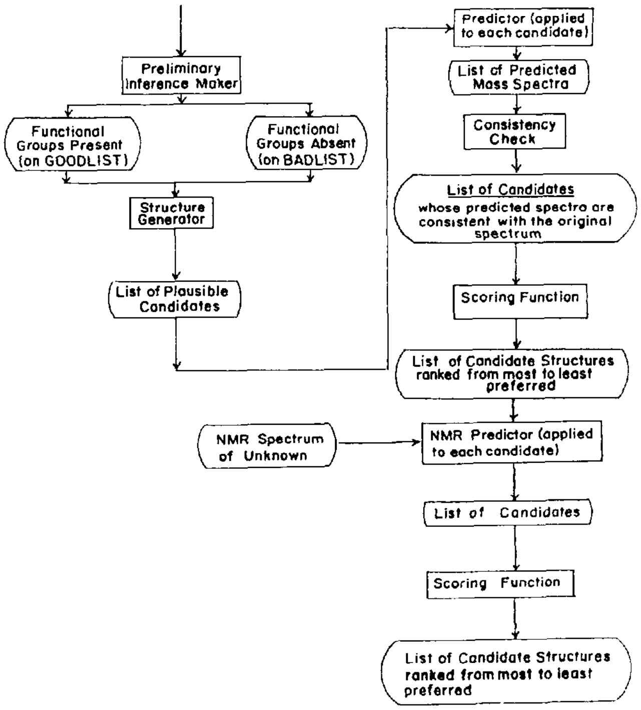


Figure 1.


Preliminary Inference Maker
Functional Groups Present (on GOODLIST)
Functional Groups Absent (on BADLIST)
Structure Generator
List of Plausible Candidates
Predictor (applied to each candidate)
List of Predicted Mass Spectra
Consistency Check
List of Candidates whose predicted spectra are consistent with the original spectrum
Scoring Function
List of Candidate Structures ranked from most to least preferred
NMR Spectrum of Unknown
NMR Predictor (applied to each candidate)
List of Candidates
Scoring Function
List of Candidate Structures ranked from most to least preferred

Scheme I. Conceptualization of Heuristic Dendral

[1] Part II: A. M. Duffield, A. V. Robertson, C. Djerassi, B. G. Buchanan, G. L. Sutherland, E. A. Feigenbaum, and J. Lederberg, *J. Amer. Chem. Soc.*, **91**, 2977 (1969).

[2] The financial assistance of the Advanced Research Projects Agency (Contract SD-183), the National Aeronautics and Space Administration (NGR-05-020-004), and the National Institutes of Health (Grants GM 11309 and AM 04257) is gratefully acknowledged.

[3] Recipient of a Fullbright Travel Award. Permanent address: Chemical Laboratory II, University of Copenhagen, Denmark.

[4] B. Pettersson and R. Ryhage, *Ark. Kemi*, **26**, 293 (1967); L. R. Crawford and J. D. Morrison, *Anal. Chem.*, **40**, 1469 (1968); **41**, 994 (1969).

[5] R. Venkataraghavan, F. W. McLafferty, and G. E. Van Lear, *Org. Mass. Spectry.*, **2**, 1 (1969); S. Sasaki, H. Abe, and T. Ouki, *Anal. Chem.*, **40**, 2220 (1968).

[6] K. Biemann and P. V. Fennessey, 14th Annual Conference on Mass Spectroscopy, Dallas, Tex., May 1966, p 322; A. Mandelbaum, P. V. Fennessey, and K. Biemann, 15th Annual Conference on Mass Spectroscopy, Denver, Colo., May 1967, p 111.

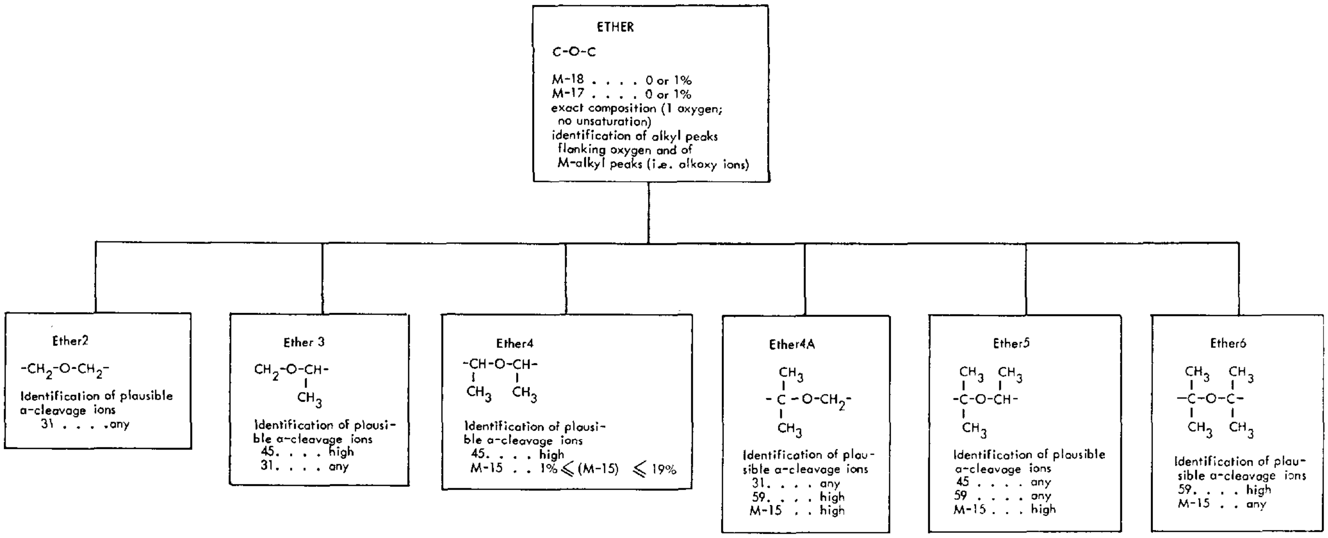
[7] J. Lederberg, G. L. Sutherland, B. G. Buchanan, E. A. Feigenbaum, A. V. Robertson, A. M. Duffield, and C. Djerassi, *J. Amer. Chem. Soc.*, **91**, 2973 (1969).

[8] H. Budzikiewicz, C. Djerassi, and D. H. Williams, "Mass Spectrometry of Organic Compounds," Holden-Day, San Francisco, Calif., 1967, Chapter 6.

[9] Program MODULES are labeled in small capitals.

[Reprinted from the Journal of the American Chemical Society, 91, 7440 (1969).]

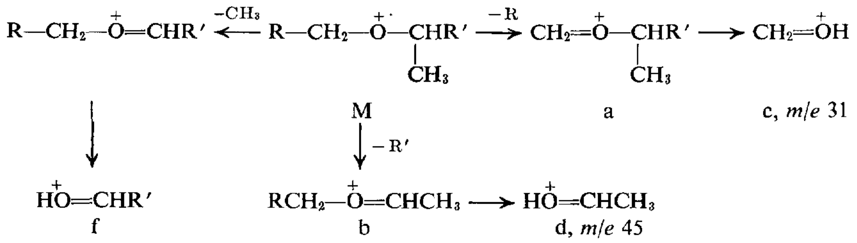
Scheme II. Rules for Ether Identification[a]


ETHER
C-O-C
M-18 . . . . 0 or 1%
M-17 . . . . 0 or 1%
exact composition (1 oxygen; no unsaturation)
identification of alkyl peaks flanking oxygen and of M-alkyl peaks (i.e. alkoxy ions)
Ether2
-CH2-O-CH2-
Identification of plausible a-cleavage ions
31 . . . . any
Ether 3
CH2-O-CH-
CH3
Identification of plausible a-cleavage ions
45. . . . high
31. . . . any
Ether4
-CH-O-CH-
CH3 CH3
Identification of plausible a-cleavage ions
45. . . . high
M-15 . . 1% ≤ (M-15) ≤ 19%
Ether4A
CH3
- C - O-CH2-
CH3
Identification of plausible a-cleavage ions
31. . . . any
59. . . . high
M-15 . . high
Ether5
CH3 CH3
-C-O-CH-
CH3
Identification of plausible a-cleavage ions
45 . . . . any
59 . . . . any
M-15 . . . high
Ether6
CH3 CH3
-C-O-C-
CH3 CH3
Identification of plausible a-cleavage ions
59. . . . high
M-15 . . any

[a] High, >10% relative abundance; any, ≥1% relative abundance.

for each possible structure is compared to the original unknown (Figure 1). Any irreconcilable difference between the unknown and predicted mass spectra results in the rejection of that candidate structure from further consideration. All the viable structures are then processed by the SCORING FUNCTION which ranks them in order of preference. At this level of the program an nmr spectrum is predicted for each surviving candidate and the results compared to the nmr spectrum of the unknown compound. In our experience this yields only one acceptable structure. The decision rules and the structure of Heuristic DENDRAL are perhaps best appreciated in a step by step discussion of its solution to a given problem.

The criteria for Heuristic DENDRAL to infer the presence of an ether function from an examination of an unknown low-resolution mass spectrum and the composition of the molecular ion are summarized in Scheme II. The program acknowledges the presence of the ether subgraph by checking for affirmative answers to the following specific points. Peaks corresponding to the loss of 17 and 18 amu, respectively, are below 2% relative abundance;[10] the empirical composition of the molecular ion must be consistent with the presence of an ether linkage within a saturated molecule and two alkyl ions corresponding to the alkyl chains flanking the ether-oxygen atom must be present. Should these conditions be satisfied then Heuristic DENDRAL attempts to expand the ether subgraph into any of the six subgraphs depicted in Scheme II. If any condition fails, none of these other ethers will be considered. The degree of substitution on either α-carbon atom will affect the masses of the products of α-cleavage of aliphatic ethers. The α-cleavage peaks referred to in Scheme II have their origin in the following mathematical relationship of the α-cleavage processes, for example, of an ether 3: M + 58 = a + b. Thus for the program to respond that an ether 3 subgraph is present it must recognize two peaks whose sum is equal to the molecular weight plus 58 amu. For ether 2, ether 4, ether 4A, ether 5, and ether 6 the masses of the radicals duplicated in α cleavage are 44, 72, 72, 86, and 100 amu, respectively. The values depicted in Scheme II as 31. . .high, 45. . . high, etc. correspond to the mass of the rearrangement ions c and d for the case of an ether 3.



The following responses were generated by Heuristic DENDRAL as it processed a typical problem. The operator initiates the program by typing the following command.[13] The program fetches the low-resolution mass

```
*(INFER (QUOTE C6H14O) S:ETH-TERT-BUT (QUOTE TEST11))
```

spectrum in question, and following an initial examination of Figure 1 by the PRELIMINARY INFERENCE MAKER the computer responds with

```
*GOODLIST = (*ETHER2!* *ETHER4A!*)

*PARTITIONS = ((*ETHER2!* 15. 43.) (*ETHER4A!* 15. 15.))
```

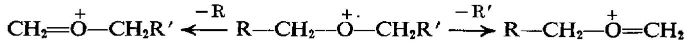
The program deduces that both the ether 2 and ether 4A subgraphs (Scheme II) are consistent with the information contained in Figure 1. (GOODLIST,[1] as the name implies, is a list of subgraphs which are thought to be

[10] The empirical composition of an ether is also compatible with the presence of an alcohol group. However, the latter class of compounds shows appreciable peaks (>2% relative abundance) in their mass spectra corresponding to the loss of water from their molecular ions.[11] Furthermore, the mass spectra of some aliphatic ethers[12] display weak peaks (<2% relative abundance) due to the expulsion of 17 amu.

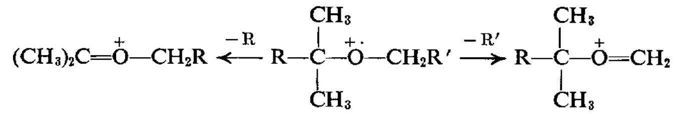
[11] Reference 8, Chapter 2.

[12] F. W. McLafferty, *Anal. Chem.*, **29**, 1782 (1957).

[13] S-ETH-TERT-BUT is the code under which the "unknown" low-resolution mass spectrum (Figure 1) is filed. It corresponds to the data recorded[12] for ethyl *t*-butyl ether and TEST 11 is the name of the storage location in which results will be kept for later use.



ether 2
(partitions R and R′)



ether 4A
(partitions R and R′)

particularly good for solving the problem at hand.) Furthermore it defines partitions which correspond to the alkyl chains expelled in the α-cleavage fragmentation of an ether 2 and ether 4A (where R and R′ are partitions for hypothetical ether 2 and ether 4A subgraphs). Finally subgraphs that appear to be poor solutions for this problem—*i.e.*, subgraphs whose conditions are violated by Figure 1—are placed on BADLIST. The alcohol subgraph is placed on BADLIST as Figure 1 contains no prominent peak due to the loss of water from the molecular ion.

```
*BADLIST = (*C-2-ALCOHOL* *PRIMARY-ALCOHOL* *ALCOHOL* *ETHER5* *ETHER
 4* *ETHER3*)
```

The command

```
*(EXPLAIN (QUOTE TEST11)(QUOTE TEST11A)(QUOTE MAR2Ø))
```
[14]

instructs that part of the program known as the STRUCTURE GENERATOR to locate the output of the PRELIMINARY INFERENCE MAKER (in file TEST 11) and the STRUCTURE GENERATOR then builds all the candidate structures consistent with the GOODLIST and BADLIST constraints, leaving the result in the external file under the label TEST 11A. The teletype response is in the following form:[15]

```
(FILE READ)
(NOVEMBER-15-1968 VERSION)
C4*ETHER2!*H1Ø
MOLECULES        NO DOUBLE BOND EQUIVS
  1.    CH2.. C3H7   O.C2H5 ,
  2.    CH2.. CH..CH3 CH3   O.C2H5 ,

(NOVEMBER-15-1968 VERSION)
C2*ETHER4A!*H6
MOLECULES        NO DOUBLE BOND EQUIVS
  1.    C.... CH3   CH3   CH3   O.C2H5 ,

DONE
*
```

The PREDICTOR section of Heuristic DENDRAL (see Scheme I) is made operational by typing the sentence

```
*(SCORE (QUOTE TEST11A) S:ETH-TERT-BUT)
```

The predicted abbreviated mass spectrum for each of the three candidate structures (read from TEST 11A) is then compared to Figure 1 to determine whether any fundamental inconsistencies exist. Those structures remaining (none were eliminated in the example under scrutiny) are then processed by the SCORING FUNCTION which ranks them in order of preference. The order depends on the number of peaks considered to be significant in the predicted mass spectrum[16] and on their estimated relative degrees of significance. For example, ions a, b, and e are assigned degree 3 and rearrangement ions c, d, and f also have degree 3. It will be observed that the SCORING FUNCTION ranks candidate 3 (ethyl *t*-butyl ether) as its first preference (score of 23). This example received an inflated score relative to the other two structures because of the branching of the *t*-butyl entity. Thus every methyl of this group is available for elimination by α cleavage (Scheme III) and each of these resulting ions can yield the rearrangement ion of mass 59. Hence, the greater the numbers of possible α cleavages the more significant peaks and the higher the score of that candidate in the present program. We have deferred the further refinement of the SCORING FUNCTION in favor of an nmr section of Heuristic DENDRAL as this promised to yield a more unambiguous result (see Figure 2).

We frequently found that mass spectrometry alone was insufficient to separate the correct structure from three or four other dialkyl ethers but that unequivocal answers could be obtained by incorporating into Heuristic DENDRAL some knowledge of nmr spectroscopy. Thus a new subroutine of the program was applied to all tenable structures passed by the SCORING FUNCTION. It should be noted that the program can profitably use nmr data if it is available, but does not require it.

The nmr program accepts two arguments, *viz.* a list of candidates (from the SCORING FUNCTION) and the nmr spectrum of the unknown compound. For each viable structure an nmr spectrum is predicted.[17] This is then compared with the unknown's nmr spectrum and the chemical shift information must agree to within ±0.3 ppm. The predicted resonance must display the same multiplicity and integral value as the unknown. If the recorded signal is a multiplet then the predicted nmr spectrum must contain one or more signals within ±0.3 ppm of this chemical shift and the values of the integrals must be compatible. If the signal requirements are not satisfied between the predicted and unknown's nmr spectrum then the disparity is noted and utilized by the NMR SCORING FUNCTION. For any candidate the score is zero if all the signals in the unknown spectrum were assigned. Otherwise the score is the product of all the integrals of the unassigned signals multiplied by 0.75

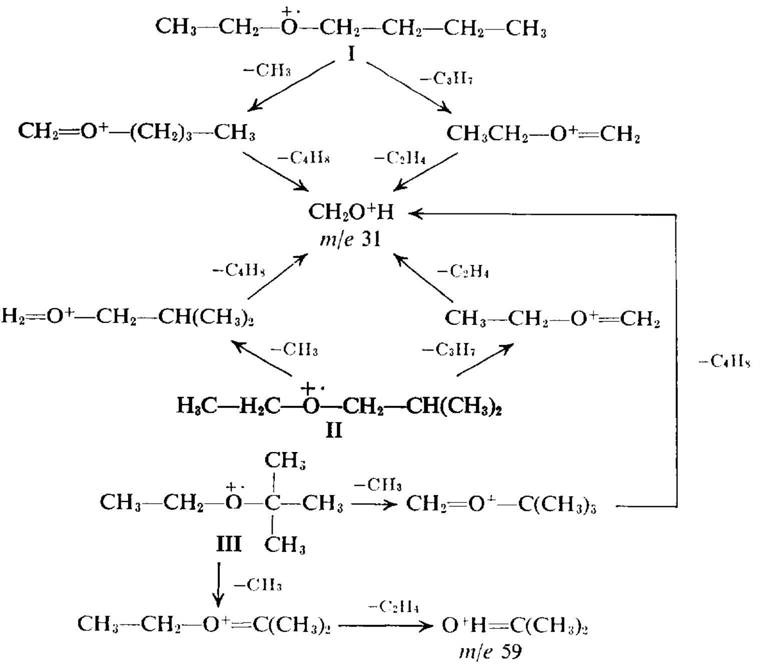
[14] "Quote" is an idiosyncracy of LISP to distinguish a label from the contents of the corresponding list.

[15] The three candidate structures represented in DENDRAL dot notation are ethyl *n*-butyl ether, ethyl isobutyl ether (both belonging to the ether 2 subgroup), and ethyl *t*-butyl ether (ether 3 subgroup), respectively. C4*ETHER2!*H10 and C2*ETHER4A!*H6 correspond to the empirical formula $C_6H_{14}O$ when the compositions (see Scheme II) C_2H_4O and C_4H_8O of an ether 2 and ether 4A, respectively, are included.

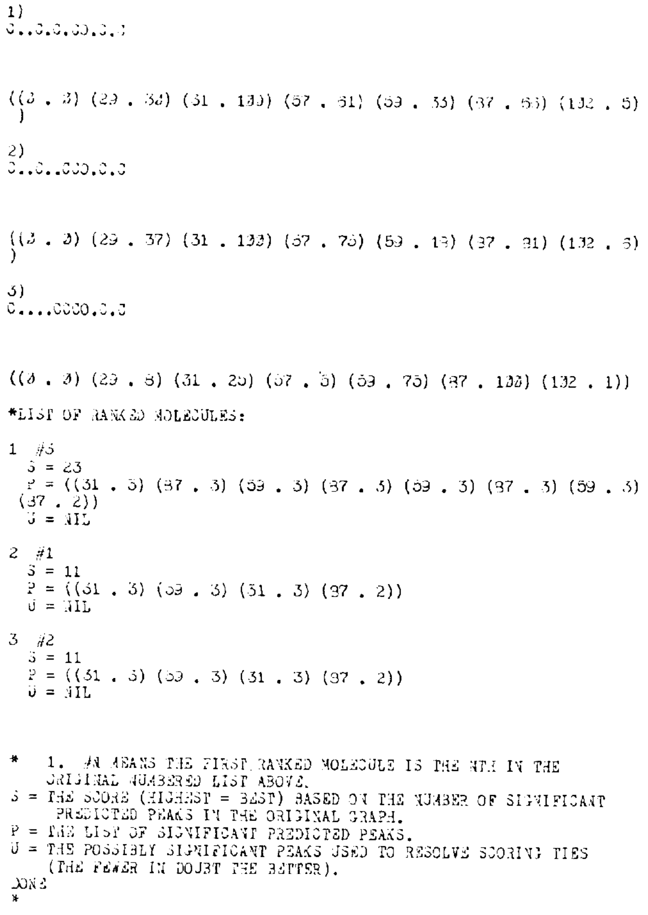
[16] In the predicted mass spectra the *m/e* value and relative intensity are listed as a dotted pair (*e.g.*, "(57.61)" refers to *m/e* 57 of 61% relative intensity). No significance should be attached to the relative intensity values as they are calculated from parameters which are at best only crude approximations.

[17] The nmr data necessary for the prediction of chemical shifts are stored as correlation tables taken from K. Nakanishi, "Infrared Absorption Spectroscopy," Holden-Day, Inc., San Francisco, Calif., 1962, p 223. The integral values for a given structure are predicted as the actual number of hydrogens giving rise to each predicted signal. The multiplicity of the predicted signal is determined by the following rules (the term "α-carbon" refers to the carbon atom adjacent to the C–H under discussion): if more than one α-carbon possesses hydrogens M (multiplet); if no α-hydrogens present S (singlet); if one α-hydrogen present D (doublet); if two α-hydrogens present T (triplet); if three α-hydrogens present Q (quartet). No use is currently made of coupling constants or other data (spin decoupling measurements) but it is anticipated that this could be incorporated into the program as required.

Scheme III



for each multiplet. The lower the score the higher the priority for any structure.



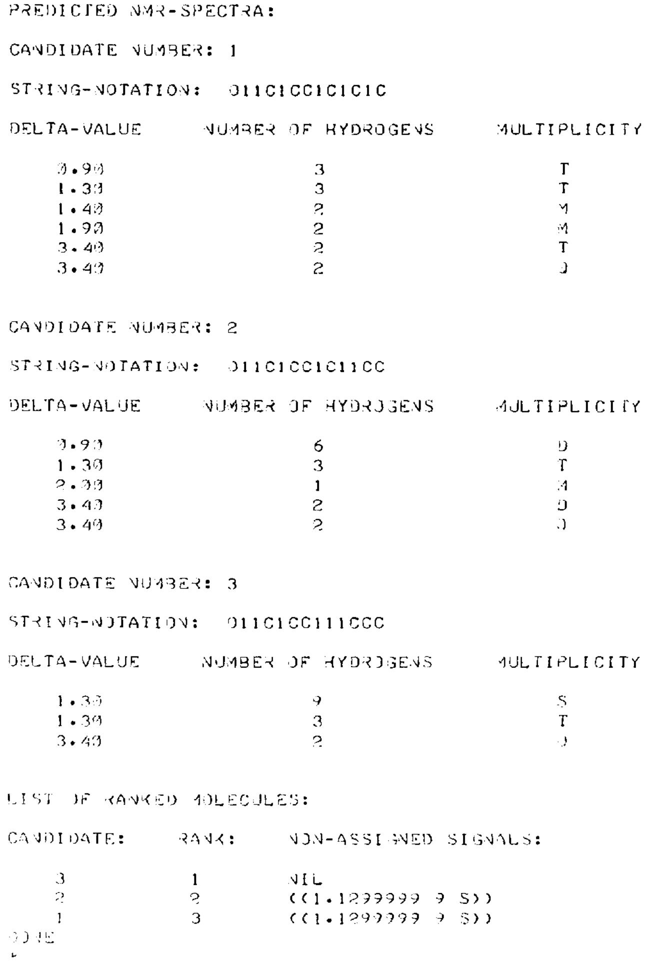
```
1)
C..C.C.CO.C.C



((0 . 0) (29 . 30) (31 . 100) (57 . 61) (59 . 55) (87 . 65) (102 . 5)
 )

2)
C..C..CCO.C.C



((0 . 0) (29 . 37) (31 . 100) (57 . 75) (59 . 19) (87 . 81) (102 . 6)
)

3)
C....CCCO.C.C



((0 . 0) (29 . 8) (31 . 25) (57 . 5) (59 . 75) (87 . 100) (102 . 1))

*LIST OF RANKED MOLECULES:

1  #3
  S = 23
  P = ((31 . 3) (87 . 3) (59 . 3) (87 . 3) (59 . 3) (87 . 3) (59 . 3)
 (87 . 2))
  U = NIL

2  #1
  S = 11
  P = ((31 . 3) (59 . 3) (31 . 3) (87 . 2))
  U = NIL

3  #2
  S = 11
  P = ((31 . 3) (59 . 3) (31 . 3) (87 . 2))
  U = NIL



*   1.  #N MEANS THE FIRST RANKED MOLECULE IS THE NTH IN THE
    ORIGINAL NUMBERED LIST ABOVE.
S = THE SCORE (HIGHEST = BEST) BASED ON THE NUMBER OF SIGNIFICANT
     PREDICTED PEAKS IN THE ORIGINAL GRAPH.
P = THE LIST OF SIGNIFICANT PREDICTED PEAKS.
U = THE POSSIBLY SIGNIFICANT PEAKS USED TO RESOLVE SCORING TIES
    (THE FEWER IN DOUBT THE BETTER).
DONE
*
```

Figure 2.

The recorded nmr spectrum (3 hydrogens, triplet at δ 1.09; 9 hydrogens, singlet at δ 1.13; and 2 hydrogens, quartet at δ 3.33) of the unknown compound (ethyl *t*-butyl ether) is already available in the literature.[18] It was presented to the program as

((1.09 3 T) (1.13 9 S) (3.33 2 Q))

and output from the program given in Figure 3 appeared at the teletype.[19]

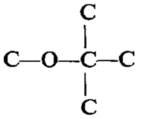
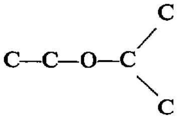
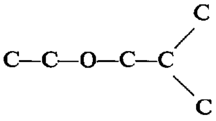
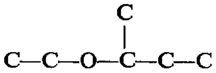
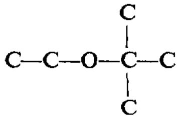
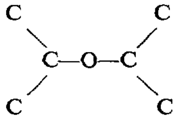
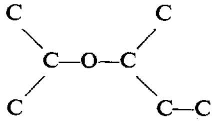
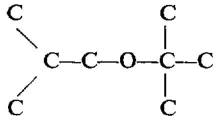
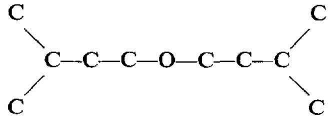


```
PREDICTED NMR-SPECTRA:

CANDIDATE NUMBER: 1

STRING-NOTATION:   011C1CC1C1C1C

DELTA-VALUE       NUMBER OF HYDROGENS       MULTIPLICITY

     0.90                3                       T
     1.30                3                       T
     1.40                2                       M
     1.90                2                       M
     3.40                2                       T
     3.40                2                       Q


CANDIDATE NUMBER: 2

STRING-NOTATION:   011C1CC1C11CC

DELTA-VALUE       NUMBER OF HYDROGENS       MULTIPLICITY

     0.90                6                       D
     1.30                3                       T
     2.00                1                       M
     3.40                2                       D
     3.40                2                       Q


CANDIDATE NUMBER: 3

STRING-NOTATION:   011C1CC111CCC

DELTA-VALUE       NUMBER OF HYDROGENS       MULTIPLICITY

     1.30                9                       S
     1.30                3                       T
     3.40                2                       Q


LIST OF RANKED MOLECULES:

CANDIDATE:     RANK:     NON-ASSIGNED SIGNALS:

    3            1         NIL
    2            2         ((1.1299999 9 S))
    1            3         ((1.1299999 9 S))
DONE
*
```

Figure 3.

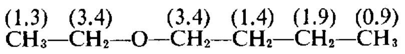
(18) H. A. Brune and D. Schulte, *Chem. Ber.*, **100**, 3438 (1967).

Table I. Heuristic DENDRAL Interpretation of the Mass Spectra[a] of Some Aliphatic Ethers

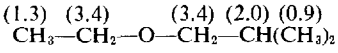
	Compound	Number of aliphatic		Number of candidates from		Ranking of candidates
		Isomers	Ethers	Structure generator	Consistency check	
1.		14	6	2	2	Correct structure ranked below ethyl *n*-propyl
2.		14	6	4	4	Correct structure ranked first
3.	C—C—O—C—C—C—C	32	15	2	2	Correct structure tied with ethyl isobutyl
4.		32	15	2	2	Correct structure tied with ethyl *n*-butyl
5.		32	15	6	6	Correct structure tied with *n*-propyl isopropyl
6.		32	15	3	3	Correct structure ranked first[b]
7.	C—C—C—O—C—C—C	32	15	1	1	Correct structure ranked first[b]
8.		32	15	10	10	Correct structure ranked first[b]
9.	C—C—C—O—C—C—C—C	72	33	2	2	Correct structure tied with *n*-propyl isobutyl
10.		72	33	1	1	Correct structure ranked first
11.	C—C—C—C—O—C—C—C—C	171	82	3	3	Correct structure tied with *n*-butyl isobutyl and diisobutyl
12.		171	82	15	15	Di-*t*-butyl ranked first Correct structure tied for second with isopropyl isoamyl
13.	C—C—O—C—C—C—C—C—C—C	405	194	17	13	Correct structure tied with 12 other ethyl ethers
14.	C—C—C—C—O—C—C—C—C—C	405	194	8	8	Correct structure tied with 7 other (C_4)—O—(C_5) ethers
15.	C—C—C—C—C—O—C—C—C—C—C	989	482	10	10	Correct structure tied with 9 others (C_5)—O—(C_5) ethers
16.		989	482	10	10	Correct structure ranked first[b]

[a] The mass spectra used as "unknown" were taken from the literature.[12] [b] Nmr spectra correctly differentiated the correct structure from the other candidates. Without the nmr input data the correct structure tied for first together with the number of candidates listed under consistency check.

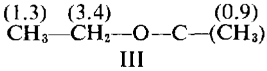
The program predicted chemical shifts for the protons of candidates 1, 2, and 3 according to the values in parentheses in structures I, II, III. Heuristic DENDRAL correctly identified the unknown from its mass and nmr spectra as ethyl *t*-butyl ether. Table I records other examples in which DENDRAL examined known spectra as "unknown" utilizing solely the mass spectral information or combining it with an nmr spectrum.



I



II



III

Although we recognize that the assignment of the correct structure to an unknown aliphatic ether is a fairly

(19) The STRING NOTATION used for candidates 1, 2, and 3 is represented in an alternative DENDRAL format in which 1 designates a single bond. These three candidates translate to I, II, and III, respectively.

simple problem it nonetheless represents a starting point for demonstrating the potential power inherent in computer interpretation of experimental data. Even when no unambiguous answers can be obtained it is impressive to note that the number of possible candidates is reduced drastically (*e.g.*, 10 candidates out of 989 theoretical possibilities in examples 15 and 16 in Table I). In the case of mass spectra taken directly from gas chromatography effluents the program would not be able to utilize nmr input data. Thus multiple solutions would be possible for a particular problem. However, as stated above, a significant degree of truncation considering all possible aliphatic ethers would be achieved. Clearly one can program other physical data (for instance ir and uv spectral parameters) to supplement the mass spectral and nmr data currently used. With added experimental data and sophisticated programming the computer should be able to solve more complex problems and it is to this end that future research in our laboratories is being directed.

Experimental Section

The computer program described here, named Heuristic DENDRAL, runs on the PDP-10 time-sharing computer at the Stanford University Artificial Intelligence Laboratory. It is written in the LISP programming language in three large parts each requiring approximately 40K of core memory (with an estimated 15K of overlap between the parts). Although many factors influence the length of time the program takes from the time it receives the initial spectrum and molecular ion composition to the time it outputs its ordered list of explanatory structures, 4 or 5 min at the teletype will usually suffice for examples of the complexity described here.

The program is now confined to monofunctional aliphatic structures. However, we are currently working on the removal of these limitations as well as adding more mass spectrometry theory to the program such that more complex problems will be within the program's capability. Details of the computer program itself have been described elsewhere.[20]

[20] B. G. Buchanan and G. L. Sutherland, "Heuristic DENDRAL: A Program for Generating Explanatory Hypotheses in Organic Chemistry," Stanford Artificial Intelligence Lab. Memo No. 62, 1968.